AD-A061 853 WASHINGTON UNIV SEATTLE DEPT OF PSYCHOLOGY F/G 5/10

AGGREGATION OF UNCERTAINTY ABOUT SUBJECTIVE JUDGMENTS.(U)

JUN 77 L C JOHNSON N00014-76-C-0193

UNCLASSIFIED TR-77-10 NL

1 OF 1

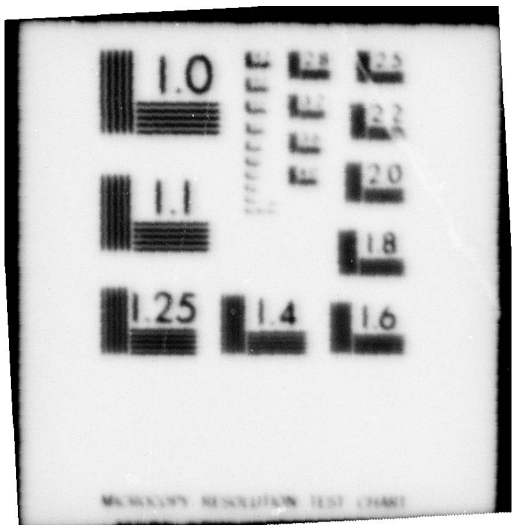
AD A061853

END

DATE FILMED

2-79

DDC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
MICROCOPY RESOLUTION TEST CHART

(12) LEVEL II

AD A061853

DECISION MAKING RESEARCH

DEPARTMENT OF PSYCHOLOGY
DEPARTMENT OF MANAGEMENT AND ORGANIZATION
UNIVERSITY OF WASHINGTON, SEATTLE, WASHINGTON

78 11 28 017

(12) LEVEL II

DECISION MAKING RESEARCH

DEPARTMENT OF PSYCHOLOGY

UNIVERSITY OF WASHINGTON

SEATTLE, WASHINGTON

(14) TR-77-10

(6)

AGGREGATION OF UNCERTAINTY ABOUT SUBJECTIVE JUDGMENTS

(10) Leonard Clark Johnson

University of Washington
Seattle, Washington

(9) Technical Report 77-10

(11) Jun[redacted]77

(12) 45p.

(15) Office of Naval Research Contract N00014-76-C-0193

(Terence R. Mitchell and Lee Roy Beach, Principal Investigators)

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: 77-10
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): Aggregation of Uncertainty about Subjective Judgments
5. TYPE OF REPORT & PERIOD COVERED: Technical Report
6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): Leonard Clark Johnson
8. CONTRACT OR GRANT NUMBER(s): N00014-76-C-0193
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Decision Making Research, Department of Psychology NI-25, University of Washington, Seattle, WA 98195
10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: NR 170-761
11. CONTROLLING OFFICE NAME AND ADDRESS: Organizational Effectiveness Research Programs, Office of Naval Research (Code 452), Arlington, VA 22217
12. REPORT DATE: June 1977
13. NUMBER OF PAGES: 45
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office):
15. SECURITY CLASS. (of this report): UNCLASSIFIED

15a. DECLASSIFICATION DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report): Approved for public release; distribution unlimited
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report):
18. SUPPLEMENTARY NOTES:
19. KEY WORDS (Continue on reverse side if necessary and identify by block number):

Aggregation of Uncertainty; Subjective Judgments; Additive Model; Averaging Model; Multidimensional; Largest Interval Strategy; Criterion Interval; Component Uncertainty; Aggregated Uncertainty

20. ABSTRACT (Continue on reverse side if necessary and identify by block number):

Four experiments were conducted to investigate the process by which humans aggregate uncertainty across tasks. The results indicate that the process can best be described as additive. However, the high variability of the data suggests both that the additive model should be regarded as only an approximation and that the process is a complex interaction of problem specific and decision-maker specific variables. As a result, the strategy by which uncertainty is aggregated is variable. Although individuals who use a heuristic are capable of consistantly applying it, those who approached the problem subjectively do not

20. show a stable pattern of responses. The complexity of the process implies that it must be appr-eched with a multidimensional perspective. The major variable clusters of the uncertainty estimation and aggregation process are specified as are their interrelationships. The implications of this research for an exter sion of MAUT techniques are discussed.

ACCESSION for

NTIS	White Section	☑
DDC	Buff Section	☐
UNANNOUNCED		☐
JUSTIFICATION		

BY
DISTRIBUTION/AVAILABILITY CODES

Dist.	AVAIL. and/or SPECIAL	
A		

Aggregation of Uncertainty about Subjective Judgments[1]

Leonard Clark Johnson

University of Washington

When one is called upon to make a subjective judgment about a person, object or event, it is common to experience some degree of uncertainty about the precision of the judgment. Judgmental uncertainty plays a large role in the psychology of decision making, and within this context the concept has received a good deal of analysis. In decision making the judgments of interest often are about proportions and probabilities of various events. In some cases the array of events is discrete (e.g., the possible diseases that a patient might have) and in other cases it is continuous (e.g., the proportion of people in the U.S. who own foreign automobiles). The work reported here deals with tasks that have continuous probability distributions.

Ignorance, risk and ambiguity are terms used to describe various states of uncertainty about population proportions or probabilities (Yates & Zukowski, 1976). If the distribution over all possible values of the proportion or probability is sharply peaked, the decision maker is fairly certain about the appropriate probability and a decision based upon it is said to be made in a state of riskiness (Luce & Raiffa, 1957). Ignorance is represented by the case in which all probabilities are seen as equally likely (Coombs, Dawes & Tversky, 1970). A rectangular distribution over the range of probabilities is the technical definition of ignorance, but from a practical standpoint, any broad, reasonably flat distribution approximates the condition. Risk and ignorance are extreme conditions which are seldom realized; all the possible variations between these extremes are said to be states of ambiguity (Ellsberg, 1961). Consequently, a continuum of uncertainty states can be envisioned in which ignorance progressively develops into ambiguity and ultimately into risk as the probability distributions become progressively more peaked. Anything that "sharpens" the distribution promotes greater certainty about the point estimate (or vice versa) and anything that "flattens" it promotes greater uncertainty.

One factor that should promote either sharpening or flattening of the distribution is the amount of pertinent information the judge has about the probability that is to be estimated. Peterson and Phillips (1966) demonstrated how observations pertinent to an event contribute to judges' knowledge about it and how subjective probability distributions (uncertainty) narrow as observations

proceed. Their subjects' task was to identify the proportion of red poker chips in a large urn on the basis of a sequence of random draws of one chip at a time. Rather than giving a point estimate of the proportion, the judges described their subjective probability distributions over the range of possible values of the proportion (.00-1.00) after each draw. This required them to give 33% credible intervals using a method that will be described below. When the obtained intervals were compared to intervals derived by using Bayes' Theorem, it was found that given enough draws, participants eventually selected a probability that was the same as that predicted by the normative model. However, they did so about half as quickly. That is, the participants were able to learn about the proportion but their speed of doing so was "conservative" compared to the normative model. Conservatism is common in probability revision experiments (Phillips, Hayes & Edwards, 1966; Peterson & Beach, 1967; Slovic & Lichtenstein, 1971) and usually is of about the same magnitude as was found in this study.

Although participants apparently could deal with it, the method Peterson and Phillips (1966) used to elicit credible intervals is quite complex:

> The Ss were told to imagine a large urn filled with poker chips, some red and the others blue. Their task was to make estimates about the proportion p of red chips in the urn when p was selected by a random procedure such that all values between 0 and 1 were equally likely. The Ss were told that they would receive information about the value of p by observing a sequence of chips drawn from the urn. Each S's task was to use two markers to trisect a scaled 0-1 continuum into three intervals such that it was equally likely that p was contained in any of the intervals.
>
> The following training procedure was used. The E displayed a practice sample of nine red chips and one blue chip. The Ss were told to use the sample as a basis for inferences about p. Each S was instructed to set his markers at 0.333 and 0.667, and then to imagine that he and two Es would each bet a dollar on which of the three intervals contained the correct proportion of red chips in the urn. Each of the bets had to be placed on a different interval and S was allowed the first choice. No S picked the 0 to 0.333 interval, and it was agreed that this interval was a bad bet. Next, S was instructed to move the markers to 0.899 and 0.901; this time it was agreed that the small 0.899 to 0.901 interval was a bad bet. The E then explained that the bets for the first settings were not equally good because too little attention had been paid to the sample of red and blue chips; too much attention was paid to the sample for the second settings. Then E instructed the Ss to set the markers so that the three intervals would be equally good bets, i.e., if the S were left with the last choice of an interval he would consider his bet to be as fair as the other two. As a second practice trial, Ss

> were told that a new urn had been selected and that a sample had been drawn which contained no red chips and six blue chips. They were again instructed to set the markers to yield three intervals, each of which would be an equally good bet. The E interrogated the Ss to insure that they understood the concept of three equally good bets.
>
> At the beginning of each sequence of 48 draws the Ss were required to set the markers at 0.333 and 0.667. They then revised and recorded these settings after observing each successive chip in the sequence of draws. The results of previous draws in the sequence were displayed throughout the sequence (p. 19-20).

Similarly complex methods have been used in other studies (Beach, 1975) and, with a good deal of practice, people often can become proficient in the use of intervals although they tend in most cases to be moderately inaccurate when compared to some objective, statistical standard (e.g., Slovic, Fischhoff & Lichtenstein, 1977).

The problem with credible intervals is that the experimenter tells the judge the criterion to use, e.g., a 33% interval, a 99% interval or whatever. While this is meaningful to the experimenter it is not necessarily so to the judge unless he is given extensive, time-consuming training. To overcome this problem Beach and Solak (1969) invented the "equivalence interval" (EI). This is the interval around some point estimate that the judge feels is "reasonably likely" to contain the true value of whatever is being estimated. Put another way, if the true value turned out to lie within the EI, the judge would count the estimate as essentially correct.

A number of studies support the contention that the EI is a useful measure of a judge's differential uncertainty about the accuracy of subjective judgments. Beach and Solak (1969) presented people with arithmetic problems (e.g., 87% of 96 = 83.5) and asked them to put EI's around the correct answers to indicate the range within which they would regard someone's answer as essentially correct if the person were to work the problem in his head. It was found that the intervals were a constant proportion, k, of the magnitude of the correct answer, C, (that is $EI/C = k$) and that k was different for difficult and easy problems.

Laestadius (1970) had people examine lists of 15 numbers of either high or low variance. They were given the correct mean of each list. Then, for each list they were asked to specify an EI around the mean within which an unaided judge's estimate of the mean would be close enough to be "in the ballpark." The EI's were significantly larger for high variance lists than for low variance lists, just as credible intervals would be.

Beach, Beach, Carter and Barclay (1974) further examined the properties of EI's. They found that the "law of proportionality" obtained by Beach and Solak (1969) held only for prothetic continua, as would be expected, but that even for methathetic continua the EI's were larger for unfamiliar events than for familiar ones. When judges set EI's around their own estimates of a population proportion based on a random sample, the EI's decreased as either the sample size increased or as the proportion approached 1.00 or .00; both conditions would affect a credible interval in the same way. Judgments of people's ages yielded k = .16 for strangers ages and .09 for strangers' judgments of the judge's own age. EI's for the seriousness of various life events (Holmes & Rahe, 1967) yielded k = .30 and for the seriousness of diseases (Wyler, Masuda & Holmes, 1968) k = .11. A final study showed that the size of the EI around hypothetical sums of money that one could inherit or give away was influenced by one's supposed wealth or poverty and by whether the money was or was not involved in a gamble. While this is a strange putpourri of topics, the studies nonetheless demonstrate that EI's vary with the variables that both common sense and statistics dictate they should.

With the exception of the proportion estimation study in Beach et al. (1974), EI's have always been placed around points that the experimenters specified rather than having the judges use them to indicate uncertainty about the accuracy of their own judgments. However, the results of the proportion estimation study showed that, when used in the latter way, the EI's behaved as credible intervals would have.

In this paper EI's will be used to examine how judges aggregate uncertainty about the accuracy of their own subjective judgments when these component judgments are compounded into overall judgments. For example, suppose a contractor were to make a series of "educated guesses" about the cost of various components of some job and then sum these guesses using pencil and paper to get an overall estimate. Each guess, however educated it may be, is still a guess and as a result has some degree of accompanying uncertainty. Therefore, the sum of the guesses also must have accompanying uncertainty. The questions of interest are: In a task like this can people aggregate uncertainty? If they can, is it possible to model the process? Does aggregation differ for easy and difficult judgments? Does the number of component judgments influence the aggregation?

Aggregation of uncertainty has been examined most thoroughly in the Bayesian revision studies (Peterson & Beach, 1967; Slovic & Lichtenstein, 1971; Slovic, Fischhoff & Lichtenstein, 1977). However, the normative model used in that research is not applicable to the present form of aggregation. Indeed, no normative model exists for this situation. Therefore we can only conjecture about possible models and then empirically seek the best.

Anderson and his associates have examined the ways in which information, as opposed to uncertainty, is aggregated in various situations. They have found that this kind of aggregation most often can be described as either additive or averaging.

Additivity means that aggregation is best described as a process in which information is merely summed as it is received. For example, a person's preference for a lunch consisting of a certain kind of sandwich and a certain kind of drink is the sum of his preference for the two separately (Shanteau & Anderson, 1969).

Averaging means that aggregation is best described as a process in which information is pooled. For example, a person's net impression of another person seems to be the average of the other's positive and negative characteristics (Anderson & Alexander, 1971).

Because these two descriptions of aggregation are simple and because they have been found adequate for a broad variety of tasks, it is reasonable, in lieu of a normative model for uncertainty aggregation, to consider them as the leading hypotheses for examining uncertainty aggregation.

The research strategy consisted of presenting participants with series (strings) of arithmetic problems and asking them to work each problem in their heads, write down the answer, and place an EI around it. Then they used pocket calculators to sum the answers to the component problems and placed an EI around that sum. Some strings were predominantly easy problems and some were predominantly difficult. For reasons that will become clear, the first experiment used strings of five component problems, the second used three, the third two and four.

Experiment 1

In the first study the sequences had five component problems. Estimated answers to these problems were aggregated on pocket calculators into either a sum or an average. EI's were obtained for the estimated answers for each component problem (EI_c) as well for the sum (EI_s) or average (EI_a). Of interest was whether EI_s and EI_a were related in some orderly way to the EI_c's of the component problems, implying that uncertainty was aggregated in some manner, and whether the additional step of dividing to obtain an average influences EI_a in any way.

Method

The method followed in this experiment will be described in some detail both because it is complicated and because the subsequent experiments were conducted in a similar way.

Materials

To provide a meaningful context for the problems each sequence was associated with a short cover story. There were four stories. One required the judge to imagine himself to be standing in a check-out line at a supermarket. He plans to pay cash and wishes to estimate the total cost of the groceries to see if he has sufficient money. This involves solving component problems such as "68 lbs. of dog food at 33¢ per pound" or "73 avocados at 87¢ each," etc. Similar stories and problems involved a contractor ("2998 electrical outlets at 99¢ each"), judgments of people's weight ("What would a 20 year old male who was 6 feet tall weigh?") and straightforward percentage problems ("What is 99% of 2998?"). Appendix A contains these stories and representative set of component problems.

A two-step pilot study was conducted to obtain a pool of difficult and easy problems. In the first, participants were asked to work approximately 50 problems in their heads and to rate the difficulty of each on a 5-point scale. Although rough, the results indicated that the per cent was the biggest determinant of difficulty, with the size of the number upon which the per cent operated being of less importance. For example, problems requiring that one take 25%, 50%, or 99% of a number were "easy" while those involving 87%, 13% or 47% were judged to be "difficult". Using these results, a large pool (approx. 300) of problems was constructed. Each problem was placed on a small slip of paper, and another group of participants was asked to sort the problems according to

whether they were difficult or easy (without actually solving the problems). The results were tested using a binomial test: Items were regarded to be difficult or easy if they were placed in one category consistently enough to reach a 0.05 significance level.

About 80 problems eventually were selected. Drawing from these, six sequences of five problems each were constructed. Each sequence was constructed so that the component problems had answers falling within one of three intervals along the number line. Two sequences each were constructed for the three intervals tested. The intervals represented progressively higher orders of magnitude (i.e., 10 to 100, 100 to 1000, and 1000 to 10,000), with the result that the experiment was able to test the effect of the size of the numbers being manipulated. Finally, the sequences were made predominantly difficult or easy (one each for each interval) by placing four difficult and one easy problem together to form a difficult sequence and four easy and one difficult problem together to form an easy sequence.

The problems were presented in booklets in which the order of sequence presentation as well as the problem order within each sequence were each independently randomized for every participant. Instructions at the end of each sequence requested either a summing or averaging of the component problem estimates. For each sequence one-half of the participants were instructed merely to sum and the other one-half were instructed to compute the average. For each participant one-half of the sequences required sums and one-half required averages.

<u>Procedure</u>

Participants were seated at desks and given pocket calculators and a booklet of experimental materials. The experimenter gave extensive instructions (Appendix B) that emphasized the purpose of the study, the meaning of EIs and how to record them, a series of practice problems to permit familiarization with the time limits placed on doing the problems and the range of difficulty of the problems. It is important to note that the calculators were only used to sum (average) the estimated answers to the component problems; they were not used on the EIs. Use of the calculators insured that the sums (averages) were mathematically accurate so that doubts about inaccurate adding (or dividing) would not contribute to the $EI_s(EI_a)$.

For each component problem in a sequence, participants were given 20 seconds to make an estimate of the answer and 30 seconds to place an EI_c

around this answer. Participants were specifically instructed to make use of the full 20 seconds and were not allowed to proceed from the estimate to the interval until the time had elapsed. Similarly, participants did not begin the next component problem until the 30 seconds allotted for the interval elapsed. After each of the 5 component problems had been completed, participants used the calculators to compute a sum or average of their estimates and then placed an EI_s or EI_a around this value.

Participants

Participants for all experiments to be described were solicited through an advertisement in the University newspaper. Both students and nonstudents were represented in the resulting pool and all were paid three dollars per hour for participation. Twenty-eight persons were involved in this study of whom six were subsequently dropped from analysis because of their apparent failure to comprehend the instructions and/or because they produced uninterpretable data (e.g., their EI's did not surround the estimate).

Results

To examine the effect of the difficulty manipulation the magnitude of the point estimate must be taken into account; previous research shows that EI's increase as the magnitude of the point estimate increases (Beach & Solak, 1969; Beach et al., 1974). To do this each EI is divided by its accompanying point estimate and the result, k, is submitted to analysis.

To test the difficulty manipulation the mean k for EI_s and EI_a was computed both for the easy and for the difficult sequences for each participant. For 19 out of 21 participants the mean k for the difficult strings was larger than for the easy strings ($p < .001$ by a sign test). The overall mean k for the difficult strings was .17 and for the easy strings it was .08. This result is congruent with those obtained by Beach and Solak (1969).

Using the ratio of EI_s to the products, sums, or averages of the EI_c for each sequence it is possible to test the adding and averaging hypotheses. Figure 1 shows the relative frequency of each value of these ratios across sequences and across participants. A ratio of 1.00 indicates that the hypothesis in question is a good fit. While neither of the hypotheses is really a very good fit, the adding hypothesis clearly is better than the averaging hypothesis. Thirty-nine per cent of the ratios lie within the .50-1.50 interval around 1.00, while for the averaging hypothesis this interval contains only 4 per cent.

Insert Figure 1 about here

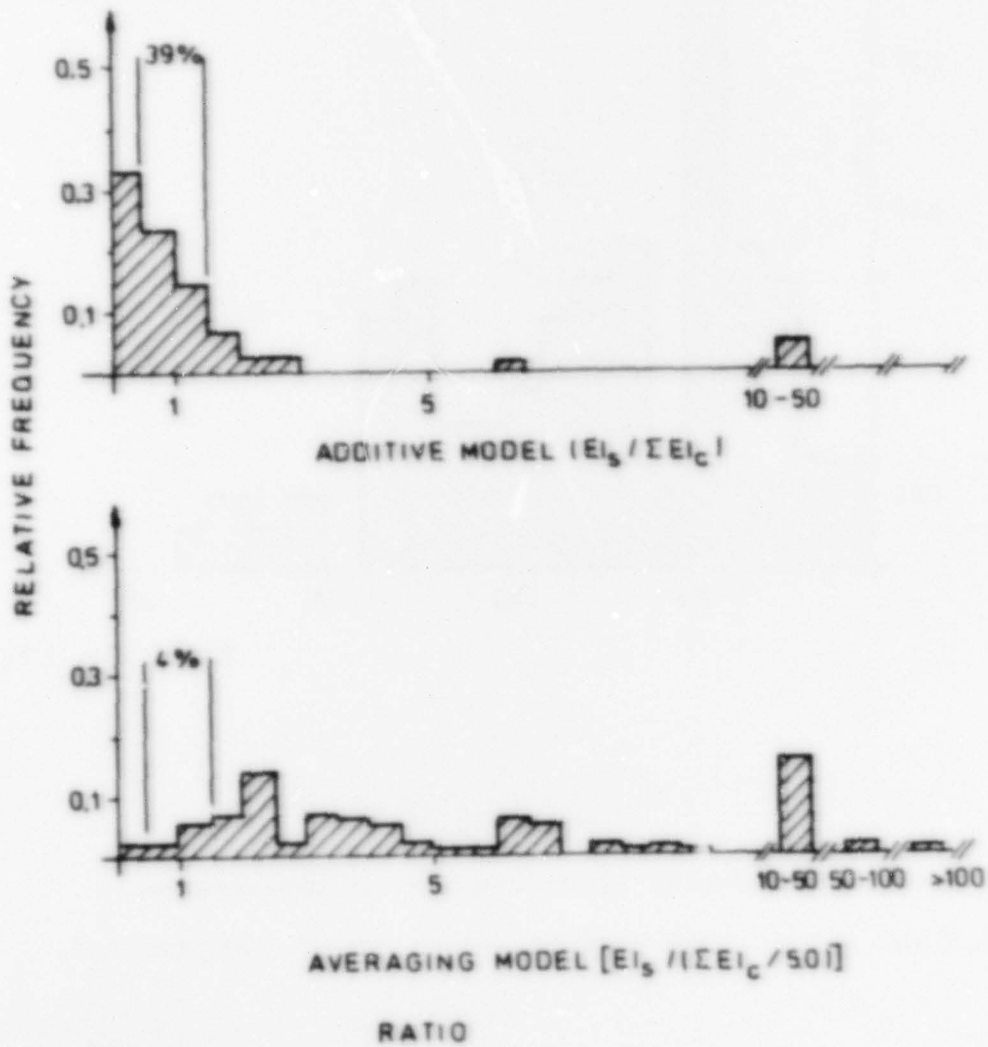


Figure 1

Comparison of Additive and Averaging Model for the Summing Sequences for Experiment 1

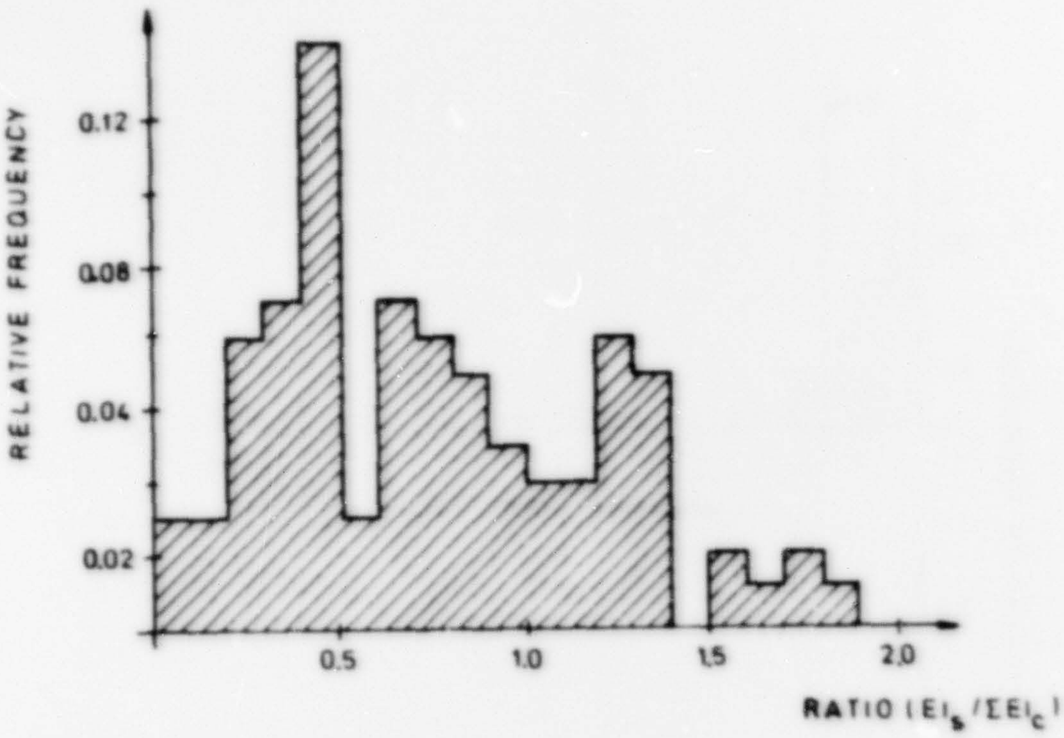


Figure 2

Relative Frequency Histogram for Ratios Between 0.0 and 2.0 for the Summing Sequences of Experiment 1

Of course, the most striking thing about these distributions is their skewness. Inspection of the raw data reveals that this results from a relatively few very extreme values of both EI_s and EI_c. These are not attributable to any particular participants, problems or sequences; they appear merely to be error. However, when used to calculate ratios these values lead to extremely large or small ratios, depending on whether it is the EI_s or the EI_c that is extreme.

Analysis of the ratios is hampered by these extreme values. Specifically, it is not possible to use standard descriptive statistics to summarize the data. In this case, the mean ratio is 2.21. The literal interpretation of this value would imply that the participants inflated the sum of the component uncertainties by a factor of roughly two. But this clearly is not the case. Figure 2 expands the additive histogram in the critical region around one. This plot further emphasizes the fact that neither the additive nor the averaging hypothesis is truly adequate. In fact, the mode (0.45) lies between the theoretically appropriate value of 1.0 for adding and 0.2 for averaging. But, in any case, the majority of responses do not lead to a ratio that is very much greater than 1.3. Figure 2 also shows just how inclusive these data are. The histograms reveal that even the "reasonable" responses form into an undifferentiated pattern in the region around and slightly below one. There could be many reasons for this, but before jumping to any premature theoretical speculations it is best to consider the simple possibility that aggregating over five component problems simply is too difficult for people to do well. To test this possibility a second experiment was conducted in which the task was made less complex.

Insert Figure 2 about here

Experiment 2

The task was simplified in two ways. First, the number of component problems was reduced from five to three. Second, participants were not asked to compute averages for any of the sequences. In addition to these changes, for the two "construction" sequences (See Appendix A) the procedure was altered so that participants worked only two problems, found the sum of these answers and placed an interval around this sum. A third problem was worked, its answer

was added to the former sum and then an EI_s was placed around this final sum. The construction problems were, therefore, a separate experimental manipulation within the body of the basic experiment. These sequences of length two were included in case strings of three component problems proved to be too difficult.

In all other respects the experiment was exactly the same as in Experiment 1.

Participants

Twenty-eight participants were drawn from the aforementioned pool of people who answered the newspaper advertisement. Six were subsequently dropped due to uninterpretable data, leaving n = 22.

Results

Repeating the previous analysis for the difficulty manipulation showed that 20 of the 22 participants had larger mean k's for difficult sequences than for eacy sequences ($p < 0.00$ by sign test). The overall mean k for difficult sequences was 0.17 and for easy was 0.06.

Figure 3 shows the relative frequencies of the values of the ratio $EI_s/\Sigma EI_c$ over problems and participants.

Insert Figure 3 about here

Comparing this histogram with that obtained in Experiment 1 it is clear that the additive hypothesis is more effective for the three component case. Specifically, the general shape of the distribution is more in line with expectations in that the median and mode are coincidental. This indicated that the obtained distribution is more stable (i.e., greater consistency in the data). Furthermore, there is a marked increase (61% vs. 39%) in the number of ratios in the 1.00 $\pm$ interval.

Although there are only two data points (two sequences) per participant, the effects of reducing the sequences to two component problems can be examined preliminarily using the sequences for which two problems were worked, their answers summed (EI_{s1}), a third problem worked, and the latter added to the first sum (EI_{s2}). The ratio of EI_{s1} and the sum of the EI_c's for the first two problems was computed for each of the two sequences for each participant. As was true for the three component sequences, 61% of the ratios for the two component sequences lie in the 1.00 $\pm$ 0.5 interval. This suggests that there is nothing gained by reducing the sequences from three to two components.

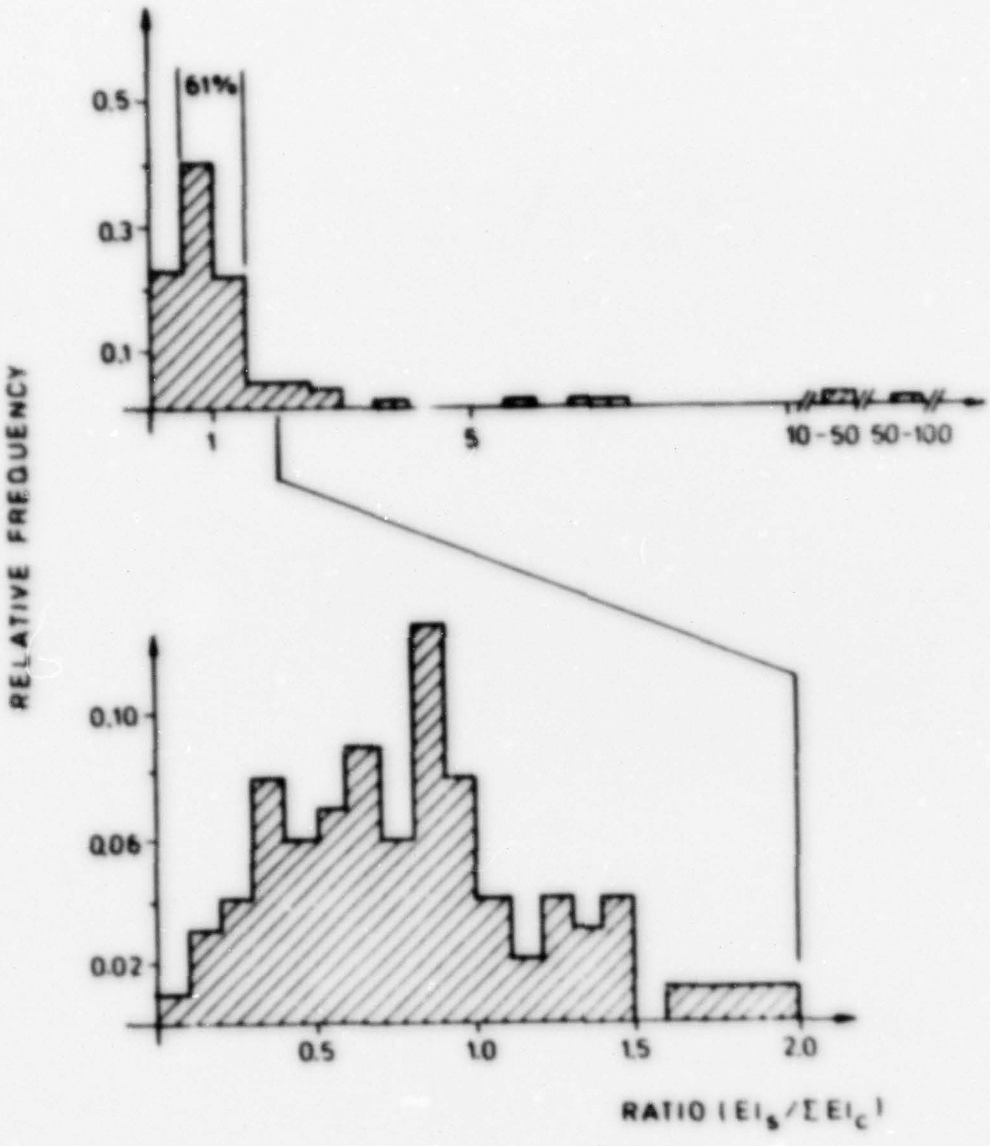


Figure 3

Relative Frequency Histogram of Ratios for the

Three Component Sequences of Experiment 2

Continuing to consider these special sequences--when the answer to the third problem is added to the sum of the answers of the first two it should be very like aggregating uncertainty for a two component sequence. In fact, this proved to be the case. Apparently, uncertainty about a sum of problem answers added to yet another uncertain problem answer is aggregated similarly to the simple two problem case.

In summary, the results of this experiment show that the responses of participants can be more effectively described by an additive hypothesis when task complexity is reduced. The improved effectiveness of the additive hypothesis in the transition from five to three components was also found for the special two and three component sequences studies in this experiment. The two component data cannot be considered as stable since there are only a very few points. Consequently, it seems reasonable to question whether two component sequences might reveal a still greater improvement in the effectiveness or an additive hypothesis (Experiment 3a).

In addition, it would be useful to obtain some idea of the effects of manipulations. The transition from Experiment 1 to Experiment 2 involved both a change in the number of components and excluded the averaging operation. It could be that for some reason participants in Experiment 1 were less able to aggregate uncertainty because of the averaging operation rather than because of the large number of components. Moreover, simple compulsivity dictates that four component sequences be examined to see if performance approximates that on five component sequences or three component sequences. Another study (3b) using four components and providing some tentative indications of the effect on the process due to a multiplying operation was undertaken.

Experiment 3a

The format for this experiment was the same as for the previous two. There were 8 sequences with 2 component problems each Sequence difficulty was manipulated by having 2 difficult items in the difficult sequences and 2 easy ones in the easy sequences. Additionally, there were sequences with a component from each difficulty level that were classed as having mixed difficulty. Of the eight sequences, four were mixed, two were easy, and two were difficult. As before, the participants worked the component problems placing EI_c around each, found a sum using the calculators, and formed an EI around this sum.

Following administration of these eight sequences there was a second experimental condition in which simple percentage problems were presented. Participants worked each problem and for each were given a number to add to their answer. An equivalence interval was then placed around the sum. Three of these simple problems were easy and three were difficult. The size of the number to be added was counterbalanced within the difficulty levels so that the three orders of magnitude discussed in Experiment 1 were approximated. Since the numerical constant had no uncertainty associated with it, this procedure can be viewed as a one-component aggregation task.

Participants

Twenty participants were obtained from the pool. Two were subsequently dropped from the analysis, leaving $n = 18$.

Results

As in previous experiments, the manipulation of uncertainty was examined using a sign test on the average string k for each participant, using only the easy and difficult sequences. For 16 of 17 participants the mean k was larger for difficult sequences than for easy ones ($p < 0.00$). The mean k for easy sequences was 0.03, for difficult sequences it was 0.19. For the mixed sequences it was 0.10.

Repeating the previously used analysis of ratios of EI_s to EI_c resulted in a mean ratio of 0.95. This is very close to the result obtained for the three component sequences in Experiment 2 as well as for the two component ($EI_{c1} + EI_{c2}$) problems. This reinforces faith in the simple adding hypothesis (see Figure 4).

Insert Figure 4 about here

When a number dictated by the experimenter is added to the answer to a single problem there should be no increase in uncertainty about the accuracy of the resultant sum. However, because this sum has been increased the EI is expected to increase along with it (Beach & Solak, 1969). Therefore, to see if the participants' actual uncertainty remained unchanged through the adding operation, the mean k for the EI_c was compared to the mean k for the EI_s. There was no difference, indicating that for single component sequences with a non-uncertain operation appended the operation does not increase relative

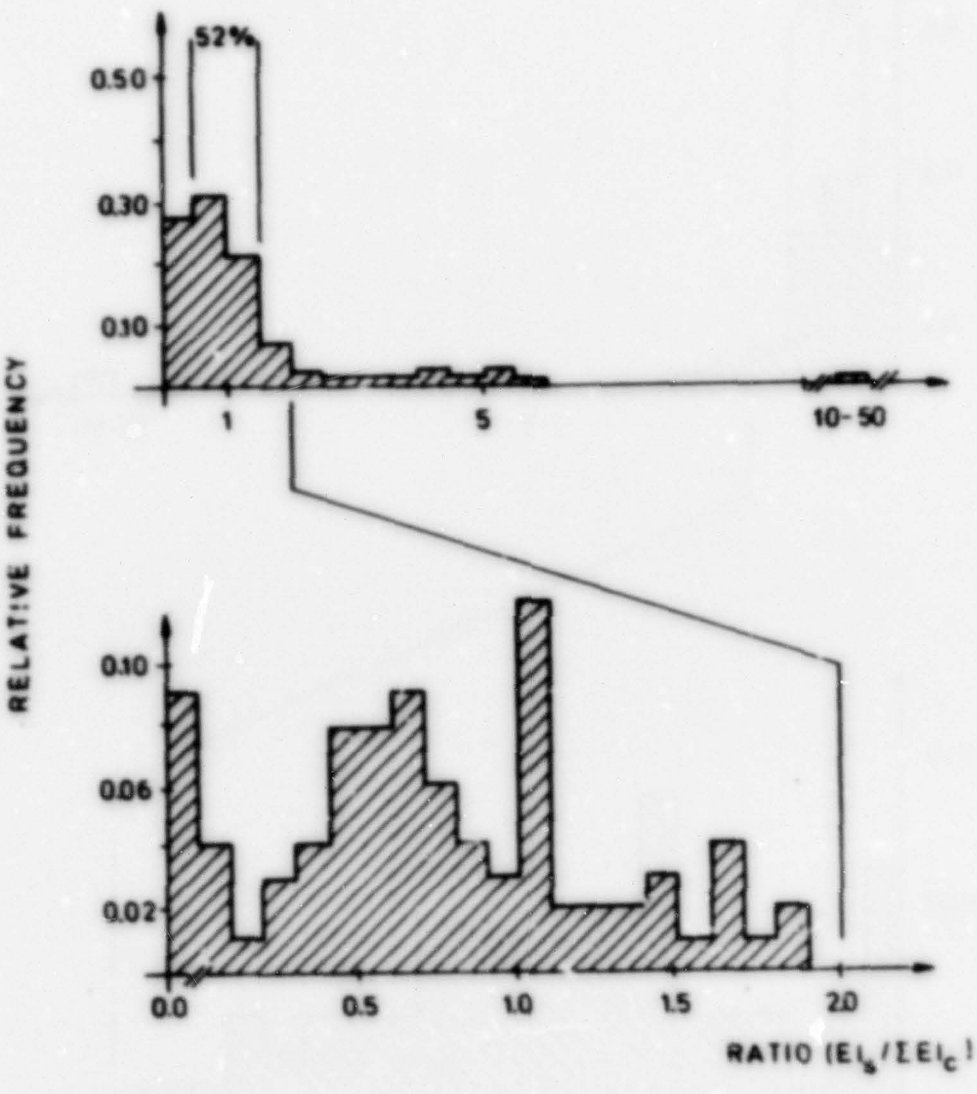


Figure 4

Relative Frequency Histogram of Ratios for the

Two Component Sequences of Experiment 3a

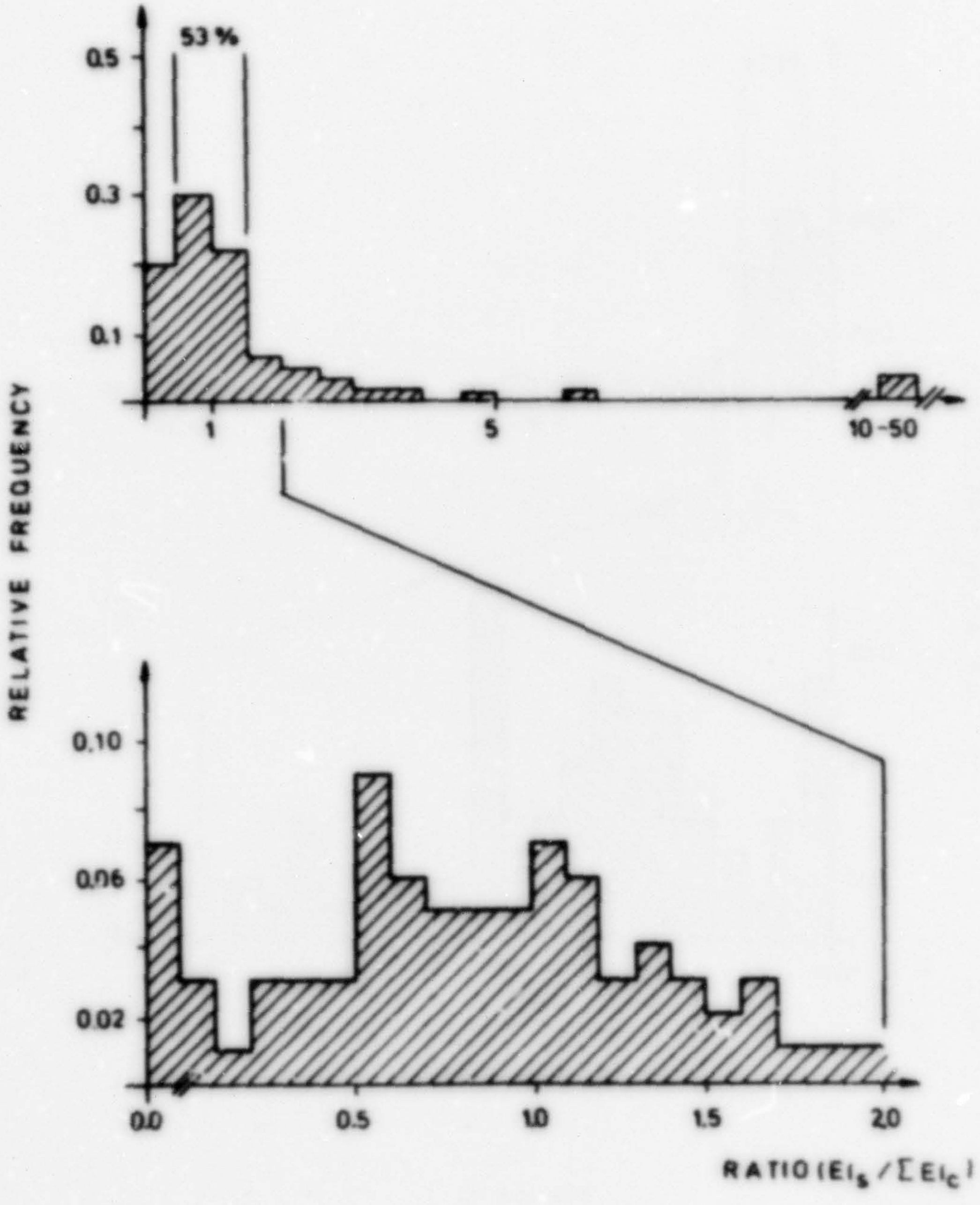


Figure 5

Relative Frequency Histogram of Ratios for the

Four Component Sequences of Experiment 3b

uncertainty. For k to remain unchanged the EI_s must have increased. Although this increase is illogical, it is consistent with the results obtained by Beach & Solak (1969).

Experiment 3b

In this study the sequences had four components and difficulty was not manipulated. Rather, the sequences were arranged so that the two sequences for each cover story were as closely matched with regard to difficulty as possible. In addition, the "grocery" problems were set up so that participants worked the sequence normally and then were required to multiply their answer by four and place an interval around this final value. The multiplication was consistent with the cover story (i.e., compute the monthly bill if the first answer were your weekly bill) and was designed to be both a test of the manipulations effect and a variation in the procedure to determine if the results in Experiment 1 could be explained on that basis. The procedure in all other respects was the same as in the previous studies.

Participants

Twenty participants were obtained from the participant pool. They were all used in the data analysis.

Results

The method used in previous experiments was applied to the data and the obtained frequency distribution is shown in Figure 5. The value of 53 per cent in the interval 1.00 ± 0.5 as well as the general shape of the curve are more similar to the two and three component experiments than to the five component case.

A comparison of the EI_s (sum EI) and the EI_m (multiplied EI) was made using a simple ratio. It seems reasonable that participants would increase their intervals by four when they multiplied; essentially paralleling their manipulation of the EI's. A ratio of 4:1 would indicate that this in fact was taking place. Using all of the data available ($N = 50$) yielded a mean ratio of 3.15. This indicates that participants did increase their intervals, but that they were unwilling to expand their intervals to the full amount that an additive hypothesis would specify as appropriate.

Insert Figure 5 about here

Although the preceding studies demonstrate the viability of an additive model for the process of uncertainty aggregation, the results are incomplete for two reasons. First, since the experiments were all conducted using students who were not necessarily skilled in the techniques of estimation, it is natural to question the generalizability of the results. "Real world experts, individuals who as a part of their professional activities make and process estimates on a daily basis, may be different from a student population that has no real interest in the task or process being studied.

Second, the experiments thus far have been designed with the hope that the data would possess low enough variability to allow the experimenter to infer a descriptive model for the process. In fact, the level of variability has proven to be quite high. Because no systematic interviews were conducted with the students, it is not possible to cross check the additive hypothesis with the subjective impressions of the participants themselves.

As a result of these difficulties a fourth experiment was undertaken.

Experiment 4

The objectives for this experiment were twofold: (1) to test the plausibility of an additive model for uncertainty aggregation using individuals with well developed estimating skills, and (2) to structure the data collection process in a way that would allow the experimenter to compare objective results of the experiment with the participant's observations of what he believed he was doing (his subjective results).

Participants

The cooperation of four practicing architects was obtained for this experiment. Although they received no financial remuneration, their enthusiasm during the experiment and their subsequent interest in the results indicates that they were highly motivated.

Materials

A floor plan of a large clinic was obtained and all measurements removed. Seven rooms were chosen from among those represented on the plan. Since the actual surface area in these rooms was held fairly constant (mean size of room was 35.6 m^2, σ = 6.4 m^2), the difficulty of estimating the surface area represented could be unambiguously manipulated by varying the complexity of the perimeter. An architect who served as an advisor to this experiment rated the chosen shapes for difficulty so that a manipulation of component problem difficulty analogous to that used in the earlier experiments could be carried out.

Procedure

The architects participated individually in a conference room provided by the firm that employed them. The experiment consisted of three phases: introduction to the task and training in EIs, the task, and a debriefing interview.

The training and introduction phase was similar to that given the student participants in previous experiments.

The task phase consisted of two trials. Each trial involved the area estimation and an EI specification for each of three rooms. Following the third room in each trial the participant was informed of the total estimated surface area of the three rooms (i.e., the sum of his three point estimates) and was asked to place an EI around this sum.

The interview phase consisted of two parts. In an initial part the participant was encouraged to describe the strategy he used in EI specification and aggregation without any specific direction being offered by the experimenter. The role of the experimenter in this section was to facilitate the discussion by restating the strategies as they were given, in order to encourage the participant to continue, and by offering comments designed to focus the participant's attention on aspects of his strategy that were not clearly described.

The second part of the interview was structured. Although an unstructured format allows participants latitude in how they describe their strategies, it often can occur that they are unable to do so with any precision. The structured format was designed therefore to give participants a series of fixed reference strategies in the hope that they would then be able to specify more precisely the differences between them and their own strategy. Moreover, there was some concern that data from the unstructured interview might not permit comparisons between individuals. By using the structured interview to provoke discussion it was hoped to obtain data that could be compared across persons.

The structured interview involved short paragraphs that were written to suggest one of three strategies: largest interval, additive, and averaging. (The paragraphs associated with each of these strategies are given in appendix C). The participants were asked to imagine that they were trying to communicate their own strategy to a person who had just expressed the

point of view represented in that paragraph. Specifically, the participants were to formulate statements which would adequately communicate their reasons for agreeing or disagreeing with the strategy portrayed.

Results

Each participant was asked to describe his method of estimating the room sizes. In every case the process of estimation was obvious to the participants. Each chose a specific feature in the plan and based upon past experience made an estimate of its size. In every case this feature was the doors which were estimated to be 1 meter wide. Using this feature as a standard, the participants attempted to estimate the length and width of the target room. Rooms that were not simple geometric shapes (i.e., rectangles or triangles) were mentally manipulated so that the surface area represented by the odd shape was translated into one of these forms and subsequently analyzed. The ease with which the participants verbalized their estimation strategy and the uniformity of this strategy suggests that the task was well suited to their area of experience and expertise.

The participants were also asked to verbalize their strategy for estimating the EI around each surface area. Here, too, the responses were strikingly consistent across all participants. Although each architect described the process in a personalized way, the main components of this description were consistent. Through discussion it became clear that the EI was perceived as a measure of their confidence in the point estimate. More detailed probing led to a list of variables that were thought to be important determinants of the EI. Specifically mentioned were perimeter complexity, experience with the type of building being discussed (i.e., hospital vs. warehouse apartment or retail store), confidence in the accuracy of the standard or modulus being used, and the extent of a person's experience in estimating. Since, for a given series of estimations, all of these factors will be constant except the complexity of the perimeter, these results suggest that the manipulation of problem difficulty was successful and unambiguous. An industry accepted heuristic of $\pm$ 10% was mentioned by 3 of the 4 participants, but none of them reported using this type of fixed k strategy; an observation that is supported by their data (Figure 6a).

Insert Figure 6 about here

		EI FOR STIMULUS 1	2	3	OVERALL RESULTANT
PARTICIPANT 1	TASK 1	5	5	10	20
	TASK 2	7	4	6	10
PARTICIPANT 2	TASK 1	3,5	4	6	16
	TASK 2	5	7	6	22
PARTICIPANT 3	TASK 1	8	8	16	40
	TASK 2	6	20	20	50
PARTICIPANT 4	TASK 1	14	8	14	13
	TASK 2	6	8	8	20

Figure 6a - Data Obtained in Experiment 4

PARTICIPANT

	1	2	3	4
TASK #1	$\sum$	$\sum +$	$\sum +$	$\bar{x}$
TASK #2	$\bar{x}$	$\sum +$	$\sum +$	$\sum +$
VERBALLY REPORTED STRATEGY	$\bar{x}$	$\sum +$	$\sum +$	WEIGHTED SUM MODEL

WHERE: $\sum$ = SUMMING

$\sum +$ = MORE THAN SUMMING

$\bar{x}$ = AVERAGING

$\bar{x}+$ = MORE THAN AVERAGING

Figure 6b - Comparison of Strategies Obtained in Interview With Those Inferred from the Data

Figure 6 gives the results obtained in experiment 4. Section a of this figure lists the numerical data by participant, task and stumulus as well as the EI specified for the total surface area for the three stimuli. Section b summarizes these numerical results by assigning an inferred or best approximation model to each trial. Also included is the participant's verbally reported strategy; his "model" for the process of uncertainty aggregation.

Discussion of the aggregation strategy, in sharp contrast to the previous two processes, was marked by substantial differences between individuals. Since the specifics of the aggregation strategies are in general different, it will be necessary to deal with the results of each individual rather than as a group.

Participant 1 was not able to vocalize his strategy immediately. However, his statements became progressively more specific as the interview progressed and these statements were all consistent with the averaging strategy which he ultimately specified as being correct. This participant felt: (1) that the errors over all the problems would tend to "average out" (i.e., an accurate estimate with relatively small EI should have as much of an effect on the aggregated EI as a poor estimate whose EI is rather large); (2) that all problems should contribute in some way to the EI of the overall estimate; and (3) that confidence affects the particular strategy chosen.

This last point was pursued in some detail and although the results are not as clear as could be desired, they do provide an interesting glimpse into this heretofore uninvestigated area. The possible strategies were seen by this participant as ranging between a purely additive and purely averaging approach. In low confidence situations the hypothesis that errors will average out is the least tenable and as a result, an additive model would be the proper choice. High confidence situations are better handled using an averaging model. However, this was judged to be true only if the criteria surrounding the problem remained approximately the same. If the expectations increase along with the person's confidence, the model of choice would again be one of additive aggregation.

Participant 2 was able to immediately verbalize his strategy as a summing model that incorporated an additional term. It was clear that this participant was using an overtly conscious approach to the task based in part on his belief that a normative model for accumulated error exists in mathematics. Accordingly, he felt that the appropriate aggregated error was slightly more

than the sum of the component EI's. Since this approach disallows the possibility of compensating errors, it is distinctively non-statistical. The data from this participant are in excellent agreement with his verbalized strategy.

This participant felt: (1) that all component problems should contribute to the overall estimated uncertainty, (2) that the averaging and largest interval strategies (Appendix C) were not cautious enough, and (3) that the choice of strategy was dependent on problem characteristics and personal expertise.

He specifically mentioned an example used to introduce the concept of an EI during the introductory phase of the experiment in which he was asked to estimate the amount of money in both his own and the experimenter's wallet as an example in which, due to his <u>lack of confidence</u> in such estimates, an averaging model would be most appropriate.

Participant 3 was also immediately able to verbalize the strategy he had used to aggregate his uncertainty estimates. Although he couched the process in terms of percentages, the result was equivalent to a summing model and, as indicated in Figure 6, this observation is in excellent agreement with his data.

Here, too, the participant's strategy was implicitly based upon an assumed theory of error accumulation which was consistently and consciously applied; in other words, a heuristic.

The interview with this participant was particularly interesting because, having specified the strategy used throughout the exercise, he proceeded to argue both for and against its merits more or less concurrently. It became clear from the ensuing dialog that: (1) He would use a model in which errors were allowed to compensate if asked to do the task again, and (2) confidence was the factor that would primarily influence his choice of strategy. Specifically, he stated that high confidence situations were amenable to a model that assumed the errors were compensating and that low confidence situations were most appropriately handled by an additive aggregation procedure.

Although participant 4 was not able to verbalize a clearly defined strategy, he did specify his approach in enough detail to permit inference of a basic model. Like the other participants, he felt that all of the component problems should contribute in some way to the overall EI. He also stated that the aggregated interval should be larger than the largest component

problem interval. These observations suggest an additive model. The participant attempted to specify a weighting strategy by which the component problems were incorporated into an EI for the sum. Perimeter complexity and the size of the area being estimated relative to the total were specified as the salient features of the weighting process but it was not possible to determine the actual equation during the interview.

During the structured interview this participant was asked to respond to several bogus strategies. He characterized the "largest interval" strategy as that of a "middle-of-the-roader." Further investigation on this point revealed that he viewed his strategy as lying between summing and averaging on a continuum that could be loosely defined as conservativeness; summing being the most conservative type of response and averaging the least conservative (i.e., most risky).

In this light it is interesting to note that each of the other participants indirectly supported this organization. Participant 1, for example, characterized the summing strategy as too conservative when compared to his averaging approach while participant 2 felt averaging was overly confident (i.e., overly liberal) when compared to his summing strategy.

In addition, the strategies used in the structured interview were consistently accepted as plausible. The participants would often respond by observing that the bogus strategy was understandable or with a phrase like, "I can see what he was doing but . . ."

The results taken as a whole support the following observations:

(1) Individuals can approach the problems of uncertainty aggregation in two distinctly different ways. The first, characterized by an easily verbalized strategy and a high level of congruence between this strategy and observable behavior gives the impression of being a heuristic. The second, as the antithesis of the first, appears to be a more subjective response to the task.

(2) The method by which subjective uncertainty assessments are combined, whether or not the individual is prone to using a heuristic approach, is not fixed.

(3) The choice of strategy is made on the basis of problem and personal characteristics which include the expectations surrounding the

solution, the confidence (skill and experience) of the problem solver, and the individual's personal response style (conservative, nonconservative).

(4) Those individuals who use a subjective approach to uncertainty aggregation are more likely to respond inconsistently to a set of apparently similar problems.

These four observations imply that the lack of consistency both between and within the participants of the first three experiments can be attributed, at least in part, to a process that is at best distinctively individualistic and at worst highly unstable.

This is not to say that the process is totally unsystematic. There appears to be general agreement that each component should contribute (i.e., be incorporated) to the overall uncertainty specification. In addition, all of the participants in this fourth experiment saw the possible strategies as in some way falling along a continuum of progressively more or less risky (or conservative) strategies.

In fact, the participants, irrespective of personal orientation, ordered the strategies similarly. Summing is seen as the most conservative with averaging the least conservative. Three out of the four participants felt that increasing confidence would lead to a strategy progressively more similar to an averaging or compensating error model.

Discussion

These experiments were undertaken to investigate the process by which uncertainty is aggregated. Taken as a whole they contribute to or expand the decision making literature in two areas: updating of subjective probabilities and the multi-attributable utility theory. The concept of uncertainty aggregation is not a completely new one. Researchers have extensively studied the process by which subjective probabilities for discrete events are updated in light of new information. The definition of uncertainty used in these studies, a subjective probability that a specific event will occur, is distinctively advantageous in that a Bayesian model is known to be normative. Unfortunately, this definition is not well suited to problems that cannot be defined in terms of the occurrence of a limited set of possible events. Furthermore, the fundamental finding of this literature, conservatism, is currently believed to be an artifact of the "book bag and poker chip" paradigm (Slovic & Lichtenstein, 1971). By selecting a less restrictive definition

of uncertainty it has been possible to investigate uncertainty aggregation in a new context. Specifically, the focus has been placed on the process by which uncertainty is aggregated across tasks. This change of focus can be viewed as an initial attempt to revitalize an area of research that has been prematurely set aside.

In their latest review of the decision literature, Slovic, Fischhoff and Lichtenstein (1977) point out that the axiomatic basis of the multi-attributable utility theory (MAUT) has developed rapidly in the last five years. These developments have led to a fairly cohesive set of axioms which, if satisfied, assure that the decomposition model implied by that set of axioms will lead to the choice of the best alternative from among a set of alternatives. The basis of this decomposition approach is a series of judgments about attributes relevant to the overall problem which are subsequently combined into an aggregated estimate of utility for the alternative. Although there has been some interest in MAUT's sensitivity to small errors in the specification of the component judgments (Fischer, 1972), the literature is surprisingly silent on the uncertainty that should be attributed to the aggregated utility. The procedure followed throughout this experimental series is a direct analogue to the MAUT technique and thus the questions of uncertainty aggregation that are examined relate directly to the development of this decision making tool.

The use of an equivalence interval as a measure of uncertainty brought with it certain methodological disadvantages. Very little is known of the relationship between EI's and common descriptors of probability distributions such as the variance. As a result, it has not been possible to use statistical concepts as a basis for a normative model. More importantly, the Bayesian model used in previous uncertainty aggregation studies could not be meaningfully applied to this type of problem. Drawing on the results of work in an analogous area of research, information integration (Anderson, 1970), two models were proposed. These were (1) an adding model, $EI_S = \sum_{i=1} EI_C$, and (2) an averaging model $EI_S = \sum_{i=1} EI_C/n$. The use of information integration concepts was appealing because of the broad range of topics that have already been successfully described using these simple algebraic models. Furthermore, a simple relationship between component problems and aggrenated uncertainty was expected. With that in mind a series of three experiments was undertaken

in which the characteristics of the component problems as well as the number of component problems were varied.

The first experiment used sequences of five component problems. As a manipulation check some sequences were difficult and some were easy; EI_s proved to be wider for the former than for the latter, indicating that participants' uncertainty was manipulated. The various hypotheses generate quite different predictions and, though noisy, the data ruled out a strict averaging model. However, the adding model's ability to account for the data was not particularly impressive. In theory this result could have occurred simply because the participants did not understand the problem. This explanation is ruled out for two reasons. First, the experimenter devoted extensive time to the explanation of the problems, as well as the techniques being used. Furthermore, the directions incorporated solving exact analogues of the problems to be used during the experiment and questions were encouraged and completely answered. Secondly, the data from each experiment were carefully scanned for response patterns that were not consistent with the directions (i.e., the EI boundaries did not enclose the point estimate). These participants consistently represented a very small percentage of the total sample. It was necessary to see if this inadequacy resulted from the participants' inability to aggregate uncertainty reliably or if it was the result of the burden of having to deal with five component problems. To examine this, a second experiment was performed using sequences that had only three component problems. As in the previous study problem difficulty was used as a manipulation check and yielded similar results. The additive model was more strongly supported in this experiment though the level of noise was still fairly high.

Subsequently, a two-part experiment was conducted. First, two component sequences were used; the results were similar to those in the three component case. In the second part of the experiment, the sequences had four components. The objectives were (1) to determine whether the additional procedure in Experiment 1 had been responsible for the relatively poor performance and (2) to find out if participants could deal as successfully with four as they did with three.

In spite of their highly variable responses, about half of the participants' responses were fairly well described by a simple additive model. Figure 7 shows

the percentage of rations that fell outside the criterion interval (1.0 + 0.5) as a function of the number of components (i.e., error percentage vs. length of sequence). There was decreasing ability to cope with the task as the number of component problems increased. Deteriorating prediction of the additive model was the most notable in the transistion between four and five component sequences. However, the results of these experiments were not completely satisfactory. Although an additive model provides some predictive power, the relatively high variability of the data suggests that the process involves more variables than had previously been expected. The fourth experiment was conducted in an attempt to reduce variability by controlling for the participants' estimating skill.

Insert Figure 7 about here

Specifically, Experiment 4 used expert estimators in a task consistent with their area of expertise. Although the results demonstrated that individuals can be stable aggregators of uncertainty, it was also apparent that this occurs only when a satisfactory heuristic for the task can be developed by the individual. This is seen as a special case of a more general subjective response process.

Figure 8 shows a process oriented summary of the uncertainty estimation and aggregation process obtained in Experiment 4. It is divided into those areas or characteristics that are objective and those that are subjective. Since the discussion that follows depends rather heavily on the figure for clarity, the reader is advised to use the diagram in conjunction with the following test. As expected, the estimation of component uncertainty is affected by the characteristics of the problem. However, the confidence of the individual also has a direct impact on the size of the interval. This level of subjective confidence has at least two primary sources of input. The first, environmental characteristics, involves external constraints such as time, the importance of the decision, limitations on the available tools for solution, and the expectations of the people surrounding the decision maker. In addition, the background of the individual directly influences his perceived confidence. All of the people interviewed saw the level of their experience and their previous successes/failures as important determinants of

their confidence. Although this diagram shows no feedback loop, it should be obvious that the decision process by which component uncertainties are specified is repeatedly applied until every problem has been analyzed.

Insert Figure 8 about here

The decision process by which uncertainty is aggregated is seen as requiring an aggregating strategy and a set of component uncertainties. Although other factors may be involved, it is believed that the fundamental determinant of the choice of aggregating strategy is a variable that is in some sense unique to every individual; his personal response to uncertainty. This variable appears to be a function not only of his level of confidence, but also of environmental characteristics. This organization of the strategy selection process is intended to emphasize its dynamic nature. Based upon personal characteristics and external demands the decision maker can be described as selecting the appropriate strategy from a set of theoretically acceptable strategies. All participants in Experiment 4 were able to accept the bogus strategies as plausible, indicating that the set of possible strategies is not limited to the one being used by an individual even if his strategy could best be described as a heuristic.

One way to interpret the consistency that emerges when a heuristic is used and the observed inconsistency of both the between and within participant data from Experiments 1, 2 and 3 is to view this section of the process as underdetermined. This suggests that given the same input on several occasions the resulting aggregating strategy cannot be expected to be the same. The existence of a heuristic may allow the individual to effectively bypass the "choice of strategy" section of this model thereby reducing this source of inconsistent responses.

As indirect support for this hypothesis the data from Experiment 3b were reanalyzed. A strategy for each string was inferred from the data so that the participant's consistency could be roughly guaged. These inferred strategies were reasonably stable for six of the 20 participants, while 14 had patterns that were inconsistent.

Experiments 1, 2, and 3 did not attempt to control for any of the subjective variables. It is therefore not surprising that the results do not

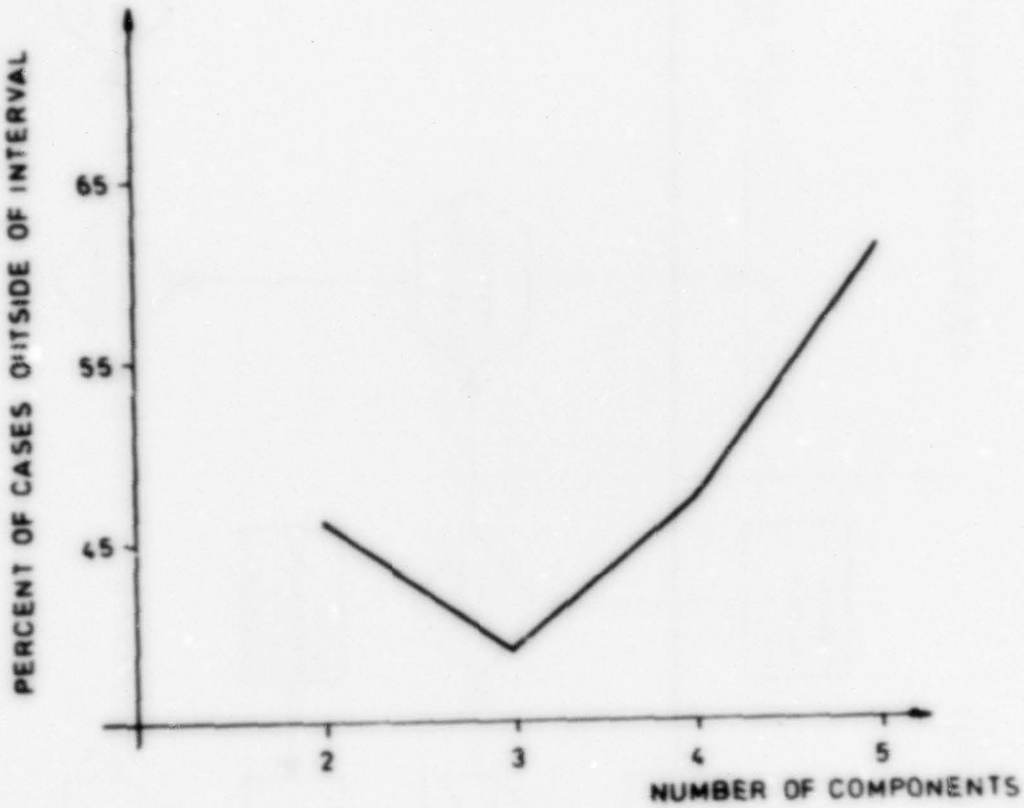


Figure 7

Per Cent of Responses Lying Outside of the Criterion Interval as a Function of the Number of Components in the Sequence

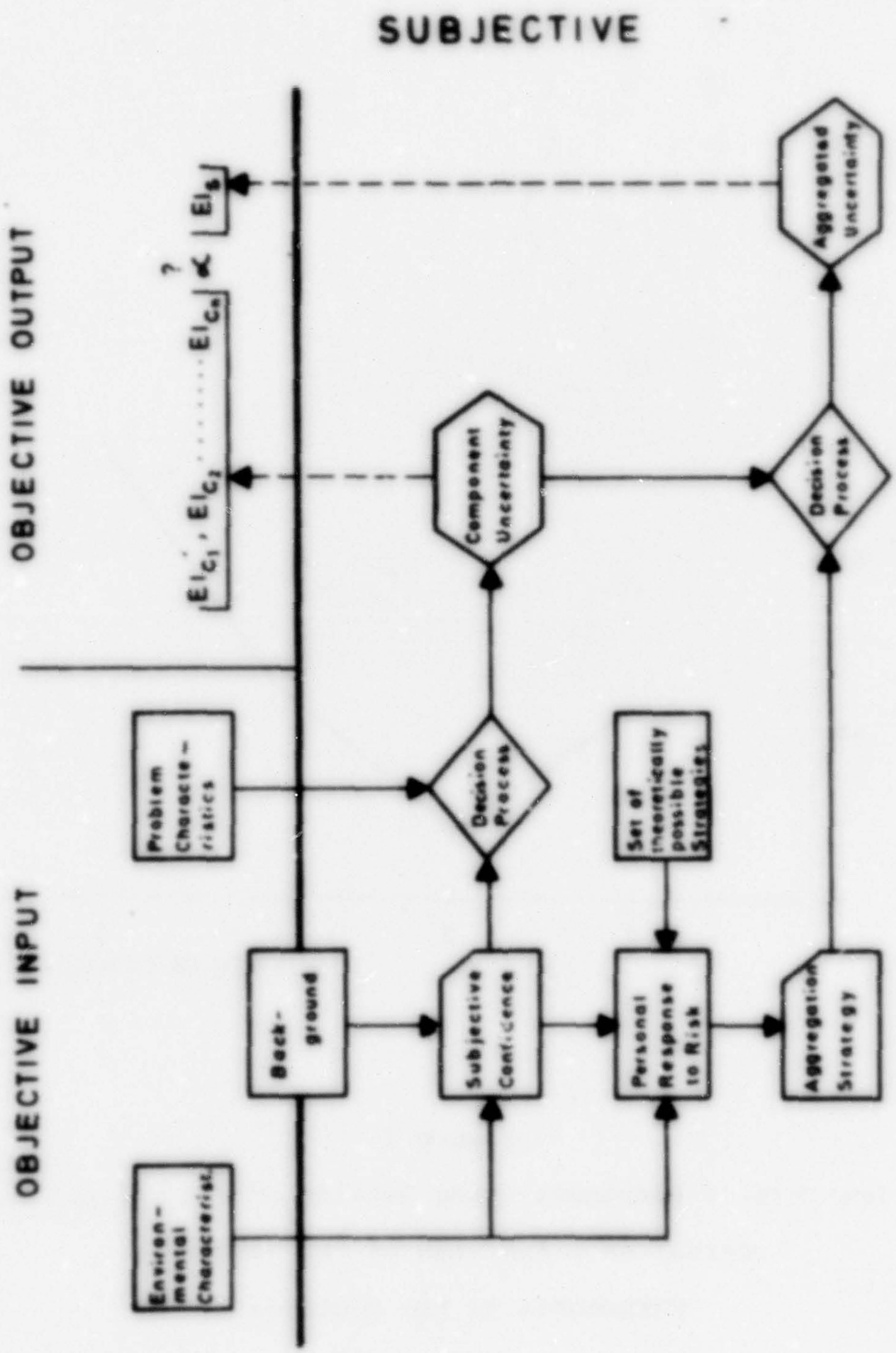


Figure 8

Summary of Component Uncertainty and Aggregated Uncertainty Process

allow a strategy to be unambiguously specified. In fact, it seems apparent that a complete specification of strategies will require a multidimensional approach that accounts both for the variables surrounding the decision (i.e., problem) and the decision maker. The work presented here suggests that man's subjective uncertainty aggregating skills are rather limited and that his choice of strategy most likely involves a complicated interaction between the objective circumstances surrounding a decision and his own internal characteristics.

It appears true that man should not be used as a standard in questions of uncertainty aggregation. Multi-attribute utility theory hinges upon the concept that man is very limited as an information integrator. This series of experiments would suggest that the uncertainty to be associated with the aggregated utility is probably best handled by a simple and defendable heuristic. One aspect of man's approach that is worthy of attention is that his response is not stereotyped. This suggests that the choice of heuristic for MAUT should be flexibly defined in a way that would allow the user to assign an uncertainty that is responsive to the needs of the situation.

References

Anderson, N. H. Functional measurement and psychophysical judgment. Psychological Review, 1970, 77, 153-170.

Anderson, N. H., & Alexander, G. R. Choice test of averaging hypothesis for information integration. Cognitive Psychology, 1971, 2, 313-324.

Beach, B. H. Expert judgment about uncertainty: Bayesian decision making in realistic settings. Organizational Behavior and Human Performance, 1975, 14, 10-59.

Beach, L. R., Beach, B. H., Carter, W. B., & Barclay, S. Five studies of subjective equivalence. Organizational Behavior and Human Performance, 1974, 12, 351-371.

Beach, L. R., & Solak, F. Subjective judgments of acceptable error. Organizational Behavior and Human Performance, 1969, 4, 242-251.

Coombs, C. H., Dawes, R., & Tversky, A. Mathematical psychology. Englewood Cliffs, N.J.: Prentice-Hall, 1970.

Ellesberg, D. Risk, ambiguity and the savage axioms. Quarterly Journal of Economics, 1961, 75, 643-669.

Fischer, G. W. Four methods for assessing multi-attribute utilities: An experimental validation (Tech. Rep. 037230-6-T). Ann Arbor: University of Michigan, Engineering Psychology Laboratory, 1972.

Holmes, T. H., & Rahe, R. H. The social readjustment scale. Journal of Psychomatic Research, 1967, 11, 213-218.

Laestadius, J. E. Tolerance for errors in intuitive mean estimations. Organizational Behavior and Human Performance, 1970, 5, 121-124.

Luce, T. D., & Raiffa, H. Games and decisions. New York: John Wiley and Sons, 1957.

Peterson, C. R., & Beach, L. R. Man as an intuitive statistician. Psychological Bulletin, 1967, 68, 29-46.

Peterson, C. R., & Phillips, L. D. Revision of continuous subjective probability distributions. IEEE Transactions on Human Factors in Electronics, 1966, HFE-7, 19-22.

Phillips, L. D., Hays, W. L., & Edwards, W. Conservatism in complex probabilistic inference. IEEE Transactions on Human Factors in Electronics, 1966, HFE-7, 7-18.

Shanteau, J. C., & Anderson, N. H. Test of a conflict model for preference judgment. Journal of Mathematical Psychology, 1969, 6, 312-325.

Slovic, P., Fischhoff, B., & Lichtenstein, S. Behavioral decision theory. Annual Review of Psychology, 1977, 28, 1-39.

Slovic, P., & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Wyler, A. R., Masuda, M., & Holmes, T. H. Seriousness of illness, rating scale. Journal of Psychosomatic Research, 1968, 11, 363,374.

Yates, F. J., & Zukowski, L. G. Characterization of ambiguity in decision making. Behavioral Science, 1976, 21, 19-25.

Footnote

[1]This research was partially supported by Office of Naval Research Contract N00014-76-C-0193 (Terence R. Mitchell and Lee Roy Beach, Investigators).

APPENDIX A

COVER STORIES

1. Percentage Problems

   The problems in your next category will be simple per cent problems.

2. Construction Problems

   In the next section the problems will be similar to those a construction estimator would be asked to solve.

   Try and place yourself in the following situation:

   You are an estimator for a construction firm. The bid is due in 20 minutes and the boss wants some figures <u>fast</u>. Glancing over your notes you come up with the following:

3. Supermarket Problems

   This time we want you to picture yourself in a supermarket.

   As you solve this set, try to keep the following scene in mind:

   You are waiting in the check-out line at your favorite supermarket. You plan to pay cash but you've gone wild on some of the specials and you may not have enough to cover the bill. A quick check of the basket reveals:

4. Weight Problems

   Americans are often concerned with their weight (overweight, underweight, ideal weights, etc.). In the next section you will have a chance to estimate some human body weights.

APPENDIX B

EXPERIMENTAL INSTRUCTION

My name is Clark Johnson. I am doing my dissertation in decision making, and this will be one of the last experiments that I am running for that dissertation. If you have read the preliminary blurb on the front page of your booklet you have a pretty good idea of what we are going to do. This really is a very straightforward experiment. It essentially involves making quantitative judgments and then supplying some additional information about these judgments.

The research is being funded by the Navy. Their primary interest in this field is in improving their ability to make distributed judgments. That is, somebody makes judgment "X" and somebody else makes judgment "Y" and somehow all of it comes together to make a decision. We are developing techniques which we can test, to see whether or not they work. This is going to be one of those techniques.

If you are a student, or if you have been a student, I am sure you have had the experience of taking an examination where the answers were a single number or some very specific thing. Sometimes you know more about what you are doing than that one number can indicate. For instance, it would be nice to be able to get more credit for the right answer (because you know for certain that it is correct, that it is the right answer) than someone who just gets that answer basically didn't have any idea in the world what they were doing. I know that has happened to me.

This whole task is quantitative in nature and there are vast differences in peoples' feelings about doing quantitative tasks. So, I have a three-point scale that I would like you to rate yourself on by writing a number between 1 and 3 on the front page of your booklet. In particular, if you are an engineer or a hard science major or an accountant, a mathematician or for whatever reasons your daily fare is numbers and you enjoy it, you would write a 3. If you are one of those people who is not particularly pleased when the end of the month rolls around and you are forced to balance your check-book, then you would use a 1. And if you are somewhere between the two extremes then you would use a 2. So, would you write something down for yourself (pause).

O.K., now as I say, this involves almost exclusively quantitative tasks, that is, manipulating numbers. One of the things that I am sure you will notice as we do them is that there are very easy problems and more difficult ones. We meant for it to be that way, and since I am not going to give you as much time as you would like to have to solve these problems, I want to encourage you not to become discouraged about not being able to answer the problems exactly. Also, I want to caution you about becoming lazy when the problems are simple. Lots of people get to a simple problem, figure "Oh boy, this is really easy," and then make big mistakes. So, please do use all of the time that I give you to actually work on solving the problem. Now, obviously, this point estimate, that single number, is not the only thing that we want. And to give you an idea of how that works we will work as a group to solve a simple judgment task and I think that that will demonstrate this simple procedure to you. Has everyone lived in this state long enough to know that there is a city called Spokane? Is there anybody who is not aware of that? O.K., why don't you write on the front page of your booklet a number that specifies your best guess as to how many miles it is from here to Spokane. We won't spend hours on this, because it is meant to be just kind of a quick little exercise. Alright, now, what are some of the numbers that we have? ("300, 200, 250"). Why don't we say 325 as a group, O.K.? I want to emphasize that it is not crucial to me what the correct answer is, this is just a question of making judgments and trying to live with them. If I came to you and said "I have an almanac here and it says that 500 miles is the distance to Spokane," if this is the true answer (500 miles), what would you say about your estimate (325 miles)? Do you think that this estimate 325 miles) is a good estimate of 500 miles if 500 is the correct answer? No, probably not. Alright, let me come down here to the other extreme and say, "What if I told you that it was 330 miles to Spokane?" Would you feel that 325 as a guess was a good enough guess of 330 as the correct answer? O.K., let's try it somewhere in between. How about 400? If 400 is the correct answer would 325 be a good guess? or a bad guess? Good? Bad? Now we are getting into kind of a grey area. And I think it is clear to you that there are certainly numbers for which your guess is a bad guess and there are certainly other numbers for which your guess is pretty good, and somewhere in between there has to be a number that is the dividing line between those two.

I should also emphasize that there are tremendous individual differences so I don't expect that you individually would all agree as to what number that should be. I expect, in fact, that there will be differences. But let us say as a group that it was 390. Maybe you feel that it should be 420. That doesn't bother me. But somewhere up here there has to be a number that is your boundary between these two regions. We can do that on the other end as well. If 100 is the right answer, 325 is probably not a good estimate. What would you say was the number such that it was the boundary between these two regions? Your own number, what is it? "270," any others? Is that a pretty good one? Alright, now what does this mean? Well, what this says essentially is, that if the true answer lies anywhere in this interval defined between 270 and 390, then it is your personal feeling that your guess of 325 was in the ballpark, it was close enough. If it is anywhere outside of that region then it would be your statement that you had missed the problem. Are there any questions about that? We call that a "ballpark estimate" because you are basically telling me what the ballpark is around your estimate. And as I say, these things can really vary. When I did it, I thought 254 was the answer. And I think 260 is my upper limit and 245 is my lower limit. Now, obviously I have a much smaller interval than you do. Perhaps that is a reflection of the fact that I drive to Spokane every break. And if I don't know what the distance to Spokane is within a pretty right space, then I would be disappointed in myself. That should indicate to you that when you know more about something then you have higher expectations. When you know something about one of these problems, it should be reflected in the width of these intervals. When you don't know anything, then you have to tell me that. Some of what I have said is summarized on the next page, so let's turn the page and I will read with you.

> The ballpark estimate describes your <u>personal</u> expectations concerning the accuracy of an estimate.
>
> The ballpark estimate is an interval around your estimated answer such that if the true answer lies outside of that interval you would feel that you had missed the problem.

The ballpark estimate can have zero width and to demonstrate that consider this problem: "50% of 402" (written on board). If I provided three blanks like this and I told you to put the estimate here, what number would you most likely choose? (201) probably, that would be my guess too, and it would be my

expectation that I wouldn't miss that. In fact, I feel that that really ought to be the correct answer. To indicate that I would put 201 as my lower boundary on the problem and 201 as my upper boundary. Clearly, I am saying, "I am absolutely certain that this is the correct answer. If it is anything different than this, I have missed the problem." Now, I expect that there will be problems for which this will be true for you. There are some very easy problems. When that is true, please do not put zeros in these blanks. It is the difference between these two numbers which refers to zero width, not the values themselves.

The last statement on your page says that the interval need not be symmetric. What that means can be demonstrated by another example. For instance, let us say that we went to the King Dome, and we knew that, for that event, 65,000 was the maximum seating capacity. If I said, "Let's estimate the number of people here." I would put up the three blanks again and you might say "63,000 people." Because of the extra information that it cannot be greater than 65,000 you have an upper boundary. You know that it cannot be greater than 65,000. So let's just say that you put that in the blank for upper boundary. However, the lower boundary is unclear. It is very hard to estimate a crowd, so you might say 45,000 for a lower limit. Now, in terms of the number line, if this is the estimated value (63,000) then you are really making a statement that looks something like this: ".________x__." In other words, your estimate is certainly not in the middle of this interval. That is fine, because what I really want to know is what <u>you</u> know about your answer. And I am trying to allow you as much flexibility to express that as I can within this paradigm. Are there any questions? Now, as I said this is a part in a series of experiments and following experimental rigor we have to control things a little bit. That means that I have to ask your cooperation in working together through these problems. It is not possible for me to just turn you loose to do them. There are two ways that I do that. One, is these little cards that I am handing out. These are called problem guides, for lack of a better term. When you start a series of problems you should place these guides such that only one problem is exposed at a time. Your card ought to have something like low, estimate, and high written on it in the places that line up with the three blanks for each problem. Test this yourself and if you don't like the way it lines up turn your card over and make your own guide. All of the answers will be written on the white booklet paper, not on these cards.

In addition to that, I am going to control how long you are able to work on a problem and in what order. In other words, I will time you for 20 seconds to solve the problem that we are working on. By solving I mean in your head. You aren't going to use the calculator and you are not to use a pencil and paper. Just look at it and do the best you can in 20 seconds. Then I will say "stop." I need an answer in every blank so I will wait for you a short time if necessary while you get something down. Once I say stop please proceed with all haste to write something down. And then I will say "Let's go on" and I will give you 30 seconds to put an interval around the point estimate. Now, it should be obvious that I am very interested in these intervals because I am giving you more time to put them down than I am to estimate the answer. And I would appreciate it if you would really think about this as your best statement about what you know. It is not crucial to me that you don't know very much. If you don't know much, that is just fine. If you know a lot that is fine, too. What I need is some kind of personal consistency; that you do express what you know. O.K., why don't we turn the page and use the problem guide to highlight just the first problem. I have a series of six problems that you can use as trial problems. We will go through them in the standard procedure I have already described and at the end of that time we will have some more instructions. So let's begin and I will give you 20 seconds to work this problem. I want the estimate and that goes in the middle column. Stop. These are practice problems and you are free to ask questions while we are working them. Although at times it seems tedious, I would appreciate it if you would follow my directions exactly. In other words, please do not proceed to write an interval around the number until I tell you to do so. I want you to spend 20 seconds making an estimate and I am going to try to force you to do that. O.K.? Now, I will give you 30 seconds to put an interval around this. Stop. I realize that this particular problem is a very easy one and that it may not take 30 seconds. In addition, I should point out that there are times when after 20 seconds you have put something down and in the remaining 30 seconds you realize that your answer is totally fallacious, that there is no way it could conceivably be correct. Well, sometimes what people have done is that they have their answer on the number line right here and in the process of the 30 seconds they realize it is wrong so they put their interval up here. Please

do change this answer if you realize that you are totally wrong. However, if we have passed a problem, if we are completely done with one and you realize that all three of your answers were in error, then leave it alone because I would rather that you concentrated completely on the problem that we are on. Let's move on and do the next problem now. I will give you 20 seconds to make an estimate. Stop. Put something down. No more thinking, just write something. It gets better as we go along. It really becomes easier. All right, now I will give you 30 seconds to put an interval around that. Stop. Next problem.

Repeat procedure for three problems.

That is half the practice problems, are there any questions about this?

Repeat for three more problems.

All right, that is basically what we are going to be doing. That is also a pretty good sampling of the difficulty range. For any of you who find this a difficult task, I have two hints. The first is one of the most important things you can do, figure out how many digits there are going to be in your answer before you get to the decimal point. Sometimes it is worth spending 15 seconds just figuring that out. Once you know that, then start working the digits and you'll find you will be more comfortable with what you are doing.

Now let's turn to the next page. What follows now is an attempt to explain to you what to expect in the experiment. The kind of procedure that we will follow. These problems always come in groups and there is always a text preceding the group. This is an example of such a text. When you get to the text you are free to read it. Then repeat so that at some point in the experiment you won't need to read it anymore, but please put yourself in this scene. When I tell you to turn the page, which is now, (this is the demonstration) you will find two problems. We solve these as we did the practice problems, in a sequence. I will then ask you to add the numbers that you have made as estimates on these calculators.

For those of you who did not bring your own calculators, I will explain how to use the ones we will provide for you. You type in the number. I think the first problem's answer is 12.07 and you type in a plus and then the second number and then you press the equals sign to get your answer. These calculators are very inexpensive and tend not to work the way you expect them

to. That is why I handed out a few extras. If you have one that has at the top a penciled-in number, like a 4 or a 2, then in the past the numbers listed have not appeared in the display after being punched. So you are reasonably certain that the calculator will work correctly if you just make sure that the number you punch does in fact appear in the display. O.K., those of you who have your own calculators have a great advantage.

Then clearly you will be adding only the estimates and that number goes down here on this line. When you do that you are free to turn the page. (We'll do that now) and on the following page there will always be the two remaining blanks for the sum that you have used the calculators to find. Now, it may not be immediately obvious, but I am sure it will seem obvious as I say it, that although you use the calculators to add these two numbers up and you know that your sum is correct for the two numbers that were in the blanks, it is not necessarily true that those numbers were correct. That is to say, you have two blanks which you add up, but the blanks were each a guess, so the true value that would appear here is not necessarily the sum of these two values. What I want to know is what is the interval that surrounds this value that makes it correct in your opinion. In other words, it is the same operation in terms of finding the intervals, but now you are going to talk about your answer as against the true sum of the two problems. Is that clear? The interval will always go on the page following the page with the problems, then there will be another little text and it starts again. This process is repeated several times. So now you should all have something talking about weight. That is the first of the sequences of problems. This completes the instructions, and if you have any questions at all, I would like you to ask them now.

APPENDIX C

PARAGRAPHS SUGGESTING: (1) LARGEST INTERVAL, (2) ADDITIVE, AND (3) AVERAGING STRATEGIES

(1) The approach I used in solving the problems was to use the biggest interval from among the component problems as my final interval. I was very careful in assigning my boundaries to begin with so I didn't see any reason to readjust them afterwards. Since the largest interval represents my weakest estimate I feel it should be used as the indicator of my overall expected accuracy.

(2) It seems obvious to me that if you add the estimates you must also add together the intervals surrounding these estimates. Since each problem has some uncertainty associated with it you have no choice but to carry all of it whether large or small into the final interval estimation. Clearly, since I couldn't use pencil and paper I wasn't able to follow this exactly but I think it accurately summarizes my basic approach to the problem.

(3) . . . Consequently, I see the process of estimating as in some way self correcting. Sometimes you're high, sometimes low, so that when you add things up you'll never be as bad as your largest interval nor as good as the smallest. I guess this boils down to saying that the final interval should be like an average of the intervals in the individual problems.

Office of Naval Research
(Code 452)
800 N. Quincy St.
Arlington, VA 22217

Director
U.S. Naval Research Laboratory
Washington, DC 20390
ATTN: Technical Info. Division

Defense Documentation Center
Building 5
Cameron Station
Alexandria, VA 22314

~~Library, Code 2029~~
~~U.S. Naval Research Laboratory~~
Washington, DC 20390

Science & Technology Division
Library of Congress
Washington, DC 20540

Psychologist
ONR Branch Office
495 Summer St.
Boston, MA 02210

Psychologist
ONR Branch Office
1030 E. Green St.
Pasadena, CA 91106

Research Psychologist
ONR Branch Office
536 S. Clark St.
Chicago, IL 60605

Director
Cybernetics Technology Office
ARPA, Room 625
1400 Wilson Blvd.
Arlington, VA 22209

Dr. H. Russell Bernard
Dept. of Sociology & Anthro.
West Virginia University
Morgantown, WV 26506

Dr. Arthur Blaiwes
Naval Training Equipment Ctr.
Orlando, FL 32813

Dr. Arie Lewin
Duke University
Duke Station
Durham, NC 27706

Dr. Lyman W. Porter
University of California
Dean, Grad. School of Admin.
Irvine, CA 92650

Dr. Paul Wall
Division of Behavioral Science
Tuskegee Institute
Tuskegee, AL 36088

Navy Personnel R & D Center
Code 01
San Diego, CA 92152

Director, Engineering Psych.
Programs, Code 455
Office of Naval Research
800 North Quincy Street
Arlington, VA 22217

Lt. Col. Henry L. Taylor, USAF
OAD(E&LS) ODDR&E
Pentagon, Rm. 3D129
Washington, DC 20301

Pers. Logistics Plans, OP987P10
Office of the Chief of Naval
Operations
Department of the Navy
Washington, DC 20350

Dr. A. L. Slafkosky
Scientific Advisor
Commandant of the Marine Corps
Code RD-1
Washington, DC 20380

~~Office of Naval Research~~
~~International Programs~~
~~Code 1021P~~
~~800 North Quincy Street~~
~~Arlington, VA 22217~~

Naval Analysis Programs, 431
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

Operations Research Prog., 434
Office of Naval Research
800 N. Quincy Street
Arlington, VA 22217

Capt. Paul Nelson
Naval Medical R&D Command
Code 44
Naval Medical Center
Bethesda, MD 20014

Director, Behavioral Sciences
Department
Naval Medical Research Inst.
Bethesda, MD 20014

Dr. George Moeller
Head, Human Factors Engin. Br.
Submarine Medical Research Lab.
Naval Submarine Base
Groton, CT 06340

Bureau of Naval Personnel
Special Asst. for Res. Liaison
PERS-OR
Washington, DC 20370

Dr. Fred Muckler
Manned Systems Design, Code 311
Navy Personnel Research and
Development Center
San Diego, CA 92152

LCDR P. M. Curran
Human Factors Engineering Br.
Crew Systems Department
Naval Air Development Ctr.
Johnsville, Warminster, PA 18974

Dr. Alfred F. Smode
Training Analysis & Eval. Group
Naval Training Equipment Ctr.
Code N-00T
Orlando, FL 32813

Dr. Gary Poock
Operations Research Department
Naval Postgraduate School
Monterey, CA 93940

Mr. J. Barber
Headquarters DA, DAPE-PBR
Washington, DC 20546

Dr. Joseph Zeidner
Dir., Org. & Sys. Res. Lab.
U.S. Army Research Institute
5001 Eisenhower Ave.
Alexandria, VA 22304

Dr. Edgar M. Johnson
Organizations & Sys. Res. Lab.
U.S. Army Research Lab.
5001 Eisenhower Ave.
Alexandria, VA 22304

Technical Director
U.S. Army Human Engineering Labs
Aberdeen Proving Ground
Aberdeen, MD 21005

U.S. Air Force Office of
Scientific Research
Life Sciences Directorate, NL
Bolling Air Force Base
Washington, DC 20332

Dr. Donald A. Topmiller
Chief, Systems Effect. Branch
Human Engineering Division
Wright Patterson AFB, OH 45433

Lt. Col. Joseph A. Birt
Human Engineering Division
Aerospace Medical Research Lab.
Wright Patterson AFB, OH 45433

Lt. Col. John Courtright
Headquarters
AMD/RDH
Brooks AFB, Texas 78235

Dr. Jessee Orlansky
Institute for Defense Analyses
400 Army-Navy Drive
Arlington, VA 22202

Journal Supp. Abstract Service
American Psychological Assoc.
1200 17th Street, NW
Washington, DC 20036

Dr. Victor Fields
Montgomery College
Dept. of Psychology
Rockville, MD 20850

Dr. Robert R. Mackie
Human Factors Research, Inc.
Santa Barbara Research Park
6780 Cortona Drive
Goleta, CA 93017

Mr. Alan J. Pesch
Eclectech Associates, Inc.
Post Office Box 179
North Stonington, CT 06359

Dr. A. I. Siegel
Applied Psychological Services
404 East Lancaster Street
Wayne, PA 19087

Dr. W. S. Vaughan
Oceanautics, Inc.
3308 Dodge Park Road
Landover, MD 20785

Director, Human Factors Wing
Defense & Civil Institute of
Environmental Medicine
P.O. Box 2000
Downsville, Toronto, Ont., CAN.

Dr. A. D. Baddeley
Director, Applied Psych. Unit
Medical Research Council
15 Chaucer Road
Cambridge, CB2 2EF ENGLAND

Prof. Dr. Carl Graf Hoyos
Institute for Psychology
Technical University
8000 Munich - Arcisstr 21
FEDERAL REPUBLIC OF GERMANY

Dr. William A. McClelland
Human Resources Research Office
300 N. Washington Street
Alexandria, VA 22314

Major David Dianich
DSMS
Building 202
Fort Belvior, VA 22060

Dr. C. Kelly
Decisions and Designs, Inc.
Suite 600
7900 Westpark Drive
McLean, VA 22101

Dr. Paul Slovic
Oregon Research Institute
Post Office Box 3196
Eugene, OR 97403

Dr. Amos Freedy
Perceptronics, Inc.
6271 Variel Avenue
Woodland Hills, CA 91364

Dr. R. A. Howard
Stanford University
Stanford, CA 94305

Dr. Ward Edwards
Director, Social Science
Research Institute
University of South. California
Los Angeles, CA 90007

Robert G. Gough, Major, USAF
Associate Professor
Department of Economics,
Geography and Management
USAF Academy, Colorado 80840

Dr. T. Owen Jacobs
P.O. Box 3122
Ft. Leavenworth, KS 66027

Dr. Delbert M. Nebeker
Department of the Navy
Navy Personnel Res. & Dev. Ctr.
San Diego, CA 92152

Professor Ken Hammond
Department of Psychology
University of Colorado
Boulder, CO 80302

Dr. Charles Gettys
Department of Psychology
University of Oklahoma
Norman, OK 73069